3.21 Machine Translation at the Document Level

Keywords

Statistical machine translation; discourse structure; connectives; pronouns; verbs

Key contact researcher

Dr. Andrei Popescu-Belis
andrei.popescu-belis@idiap.ch
Tel.: +41 27 721 77 29

Technology Transfer Office

Dr. Florent Monay
Dr. Hugues Salamin
tto@idiap.ch
Tel.: +41 27 721 77 72

Corporate Sponsorship Program

See Section 4 of the present document

File reference & version number:

Software disclosure 8452
Software disclosure 9011

Functional description

This technology improves current systems for automatic translation, by using text-level information. Several pre–processors detect inter-sentence dependencies and use them to label different types of words, prior to translation. The labels are used by a third-party statistical phrase-based machine translation system (open source) which otherwise translates sentence-by-sentence. We currently deal with discourse connectives (e.g. 'since' or 'while'), pronouns ('it' vs. 'il' or 'elle') and verb tenses in English, and with Chinese and German compounds. Our method improves the translation of entire texts.

Innovative aspects

- Coupling of discourse-level classifiers with SMT
- Learning to detect semantic features for connectives, pronouns, compounds, and verbs from parallel corpora

Commercial application examples

- Improved MT systems for long coherent texts
- Post-editing tools for existing MT engines

More information

Thomas Meyer, Najeh Hajlaoui, and Andrei Popescu-Belis, "Disambiguating Discourse Connectives for Statistical Machine Translation", *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 23(7):1184–1197, 2015.

Software & IPR status

- Open source software for translation of connectives https://github.com/idiap/DiscoConn-Classifier
- and their evaluation https://github.com/idiap/act

Research software and know-how for other types of words (verbs, pronouns, compounds).